A Distance Learning Testbed

William L. Bewley, Gregory K. W. K. Chung, Jin-Ok Kim, John J. Lee, and Farzad Saadat
UCLA/CRESST
Los Angeles, CA
{ bewley, saadat }@cse.ucla.edu, {greg, jinok, johnjn }@ucla.edu

ABSTRACT

Because of the great promise of distance learning for delivering cost-effective instruction, there is great interest in determining whether or not it actually is effective, and—more interesting—determining what variables of design and implementation make it more or less effective. Unfortunately, much of the research has been based on simple comparisons of distance learning to the "traditional" method of instruction rather than examining the variables influencing the effectiveness of distance learning. In addition to not manipulating or controlling important independent variables, the dependent measures used in such studies are often inappropriate, ranging from the obviously inadequate, e.g., the "smile test," to standardized tests that have known psychometric properties but are not aligned with course objectives, to homegrown measures that appear to be aligned with instructional objectives but are of unknown reliability and validity. We have addressed the problem of limitations in dependent measures with research on measures of student achievement based on families of cognitive demands, and have developed assessment models for these families that can be used to design assessments across a variety of subject matters. We have also developed computer-based assessment tools implementing the models, including tools for data collection, scoring, analysis and reporting, assessment authoring, and knowledge acquisition and representation, and with support from the Office of Naval Research we have developed a distance learning testbed to apply these models and tools to distance learning research and evaluation. This paper describes our current (summer of 2004) testbed implementation and presents three examples of the research conducted in the testbed on methods for assessing human performance via distance learning technologies.

ABOUT THE AUTHORS

William L. Bewley is an Assistant Director of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). His experience combines research on human learning and performance, instructional systems design, education and training product development, software development, and management. His current work is concerned with research on applications of advanced technology to human performance assessment in distance learning and team problem-solving environments and evaluation models for e-learning. He has advised the National Science Foundation, commercial companies, and non-profit organizations on educational technology and integration of technology into curricula. Dr. Bewley earned a Ph.D. and M.S. in Psychology and Computer Science from the University of Wisconsin and a B.A. in Psychology from Stanford University.

Gregory K. W. K. Chung is a Senior Research Associate at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). His current work at CRESST involves developing computer-based assessments to measure problem-solving and content knowledge in military and engineering domains. He has experience developing Web-based assessment tools for diagnostic and embedded assessment purposes using Bayesian networks, constraint processing, domain ontologies, and other advanced computational tools. Dr. Chung earned a Ph.D. in Educational Psychology from the University of California at Los Angeles, an M.S. degree in Educational Technology from Pepperdine University at Los Angeles, and a B.S. degree in Electrical Engineering from the University of Hawaii at Manoa.

Jin-Ok Kim is a Ph.D. candidate in Social Research Methodology at the University of California at Los Angeles and a Research Assistant at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Her work is concerned with the combination of causal inference and use of statistical methodologies, especially hierarchical modeling, in various empirical studies (e.g., longitudinal analysis, program evaluations, and field experiments). She earned a B.A. in Education from the Seoul National University.

John J. Lee is currently a Senior Research Associate at CRESST working on the development of the Quality School Portfolio (QSP) web-based decision support software for school districts. He is also involved in the development of the Assessment Design and Delivery System (ADDS), an online tool to help teachers create more challenging assessments, including knowledge mapping, explanation, and simulation tasks. He received his Ph.D. in Psychological Studies in Education at UCLA in the Graduate School of Education and Information Studies, a Masters of Divinity from Princeton Theological Seminary, and a Bachelor of Science in Biology from the University of Michigan, Ann Arbor. His research interests include data-informed decision making, online assessment development, knowledge mapping, and assessment accommodations for English Language Learners.

Farzad Saadat is a Programmer analyst and software developer at the National Center for Research on Evaluation, Standards, and Student Testing. He works primarily in the area developing algorithms and software to help with computer-based assessments projects. He got his B.S. degree in Electrical Engineering from the University of California at Los Angeles, and an M.S. degree in Electrical Engineering/Computer Networks from University of Southern California.

A Distance Learning Testbed

William L. Bewley, Gregory K. W. K. Chung, Jin-Ok Kim, John J. Lee, and Farzad Saadat
UCLA/CRESST
Los Angeles, CA
{ bewley, saadat }@cse.ucla.edu, {greg, jinok, johnjn }@ucla.edu

Because of the promise of distance learning for delivering cost-effective anytime/anywhere instruction, there is great interest in determining whether or not it actually is effective, and—more interesting—determining what variables of design and implementation make it more or less effective. Unfortunately most of the research has been based on simple comparisons of distance learning to the "traditional" method of instruction, which usually means classroom instruction based on lecture presentation of content, rather than examining the variables that influence the effectiveness of distance learning for different people learning different skills in different environments. Like the media comparison studies of decades before (Army Science Board, 1997; Barry & Runyan, 1995; Clark, 1983, 1989, 2001; Director of Naval Training, 1998; Lockee, Burton, & Cross, 1999; Machtmes & Asher, 2000; Madden, 1998; Phipps & Merisotis, 1999; Russell, 1999; Saba, 2000; Smith & Dillon, 1999; Wisher et al., 1999), these distance learning studies have usually found no significant differences, which is not surprising because, like the media comparison studies, they fail to examine or even control for the important variables influencing learning. As Richard Clark has pointed out, media do not cause learning any more than the truck that delivers groceries causes changes in our nutrition (Clark, 1983, 2001). What's important are the variables influencing learning, e.g., the instructional strategies enabled or enhanced by the use of media.

In addition to not manipulating or controlling important independent variables, the dependent measures used in such studies are usually inappropriate, ranging from the obviously inadequate, e.g., the "smile test" that asks students if they like the course, to standardized tests that have known psychometric properties but are not aligned with the objectives of the course, to homegrown measures that appear to be aligned with instructional objectives but are of unknown reliability and validity (Baker & Herman, 2003).

The National Center for Research on Evaluation, Standards, and Student Testing (CRESST) has addressed the problem of limitations in dependent measures with research on measures of student achievement based on families of cognitive demands (content understanding, problem solving, teamwork and collaboration, metacognition, and communication) and has developed assessment models for these families that can be used to design assessments across a variety of subject matters (e.g., Baker, 1996; Baker, Freeman, & Clayton, 1991; Baker et al., 1996). CRESST has also developed computer-based assessment tools implementing these models in content understanding and problem solving (e.g., Chung, O'Neil, & Herl, 1999; Herl, O'Neil, Chung, & Schacter, 1999; Schacter, Herl, Chung, Dennis, & O'Neil, 1999; Chung et al., 2003; Delacruz, Chung, & Bewley, 2003).

Over the past three years, we have been developing a distance learning testbed to apply these models and tools to research on the variables influencing distance learning effectiveness. The testbed has been used to conduct usability studies of courseware and online assessment tools, evaluations of the effectiveness of distance learning courseware, and basic research on feedback variables influencing the effectiveness of online courseware. This paper describes our current (summer of 2004) testbed implementation and presents three examples of the research conducted in the testbed on methods for assessing human performance via distance learning technologies.

CURRENT IMPLEMENTATION

There are seven major testbed components: data collection platforms, special data collection tools, data analysis and reporting tools, a courseware rating tool, assessment authoring tools, performance assessment scoring tools, and knowledge acquisition and representation tools. The need for data collection, analysis, and reporting is obvious. Perhaps less obvious is the need for a courseware rating tool, but evaluation against a rubric derived from research-based knowledge of what works in distance learning is an important predictor of courseware effectiveness. Assessment authoring and scoring tools provide support for the difficult tasks of developing measures

appropriate to the learning tasks and then scoring the results of those assessments. Finally, capabilities for assessment authoring and scoring depend on the existence of accurate, complete, inspectable, shareable, and maintainable knowledge of domain content, and that requires support for knowledge acquisition and its representation in a usable form.

Data Collection Platforms

The primary data collection machines are 30 Dell Latitude D600 PCs, with a Pentium® M Processor 1.40 GHz, 256 MB memory, two 40 GB hard drives, a 24X maximum/10X minimum CD-ROM drive, and Intel® PRO/Wireless 2100 WLAN (802.11b, 11MBPS) wireless local area networking. The operating system is Microsoft® Windows® XP Professional. Secondary test machines are Apple® Power Mac G4s®, with 1.25GHz, 256 MB memory, an 80 GB hard drive, a combination DVD-ROM/CD-RW drive, and Ethernet. The operating system is Mac OS X 10.3. The Internet browser is Microsoft® Internet Explorer® 6.0, updated as required to remain current. Netscape 7.0 is an alternative browser for both primary and secondary test machines to support compatibility testing including interactions of browser and platform.

Wireless networking enables easy and flexible configuration of machines in our laboratory and remote sites. Special shipment cases have made it easy to transport test machines to remote sites. Much of our data collection over the past two years has, in fact, been conducted at the learners' work environment.

Special Data Collection Tools

One of the most important steps in conducting human performance research is specification of behavioral demonstrations of desirable or essential elements of competence and the collection of data from the demonstrations. Appropriate behavioral demonstrations depend on the characteristics of the competence, especially the cognitive demands. Performance requiring content understanding may, for example, require that the learner indicate key concepts and the relations among concepts by creating a knowledge map using the CRESST Human Performance Knowledge Mapping Tool (HPKMT). HPKMT functionality is described in detail in Chung et al. (2003). The research reported in this paper uses the HPKMT to measure content understanding. Performance requiring problem solving may require that the learner solve problems in a simulated environment, and data must be collected on critical aspects of the targeted behavior in that environment. For example, what does the clickstream reveal about how the learner navigates through the simulated environment, accessing information moving from screen to screen? Where on the screen is the learner is looking at particular points in the behavioral demonstration? How does the learner describe the thinking behind actions, and how does the description correlate with the clickstream and where the learner is looking? To collect these data, the Distance Learning Testbed provides special data collection tools including clickstream recording, eye tracking, audio and videotaping, and data fusion software tools that correlate the clickstream, eye tracking, and audio and videotaped data.

Data Analysis and Reporting Tools

Data from all data collection tools are recorded in a relational database. Data reduction and analysis are accomplished using commercial off-the-shelf (COTS) packages such as SPSS. In addition to these packages, CRESST has developed a Web-based data analysis and reporting tool called the Quality School Portfolio (QSP). QSP provides a framework that allows users to quickly and easily investigate relationships among variables. Learner performance data can be stored in a digital portfolio as text, images, audio, or video. Data can be disaggregated by various groups and, through a report function, transformed into easy-to-understand graphical reports.

Courseware Rating Tool

In recent work funded by the Office of Naval Research ("Knowledge, Models and Tools to Improve the Effectiveness of Naval Distance Learning"), the University of Southern California Rossier School of Education (USC RSOE) has produced research-based guidelines for courseware development and evaluation directed at five critical dimensions of a distance learning instruction and assessment system: management strategies, learning strategies (self-regulation and motivation), instructional strategies, multimedia strategies, and assessment strategies. These guidelines were documented as a set of books organized by guideline area and containing the guidelines and all cited references, and they have been posted on the Advanced Distributed Learning Initiative Web site.[1] CRESST has applied these guidelines to evaluation of distance learning courseware for the Navy civilian workforce under the ONR-funded

[1] http://www.adlnet.org/index.cfm?fuseaction=DLGuid

project "Research-Based Distance Learning to Support Navy Workplace Education." This work transformed the guidelines into a framework of questions that can be answered by entering a rating on a scale. The framework was tested for usability and rater reliability and has been implemented as a Web-based tool making it possible to apply the guidelines to distance learning courses as they are being reviewed or evaluated online, with links to worked examples and supporting literature.

Assessment Authoring Tools

CRESST's vision for assessment authoring is to create a suite of tools that will allow distance learning researchers and curriculum developers to easily create measures based on the CRESST assessment models. The first Testbed authoring tool supports creation and maintenance of CRESST HPKMT knowledge mapping tasks. Each knowledge mapping task has properties, including the set of concepts and links, the set of icons used, and the mode of operation (e.g., select-only, type-in, or both). Tasks can be composed of existing concepts and links, or the user can create new ones or can edit an existing set.

Automated Performance Assessment Scoring Tools

Because of the complexity of the data and the scoring, automated scoring is an important requirement for performance assessment systems. The testbed includes scoring tools designed to score HPKMT data. The tools are loosely coupled to the HPKMT in the sense that HPKMT knowledge mapping data are stored in tables and databases that are independently accessed by the scoring engine. The separation reduces the complexity of the mapping code, enhances maintainability, and allows scoring algorithms to be developed independently of the mapping tool. CRESST has implemented two scoring routines for the knowledge mapping representation: (a) exact proposition matching, and (b) synonym-based proposition matching. Exact-proposition-matching is based on the algorithm developed by Herl, Baker, and Niemi (1996), and involves counting the number of propositions in the learner map that also exist in the referent map, e.g., an expert's map. Synonym scoring is more lenient, counting a match if a term matches one of a set of synonyms for each term in the learner map and each corresponding term in the referent map.

Knowledge Acquisition and Representation Tools

A major goal in designing and developing the CRESST Distance Learning Testbed has been to ensure that measures meet the technical requirements appropriate to their purposes, and this goal has been supported by our assessment models and the development of tools to support use of the models. The models are based on identification of the cognitive demands associated with performance objectives and specification of behavioral demonstrations of mastery. They are also heavily dependent on analysis of content to be addressed. A model of the content domain is a critical requirement for designing assessments and scoring learner performance, and CRESST has developed knowledge acquisition and representation tools to support it.

Our approach is based on the use of a domain ontology to represent knowledge and instructional content linked to specific knowledge elements. An ontology provides a domain model, a representation of the knowledge in a domain, usually represented as a knowledge map consisting of a network of concept nodes and links connecting concepts. The ontology can be communicated to people and computational systems such as the testbed authoring and scoring tools (Fensel, Hendler, Lieberman, & Wahlster, 2003). COTS and GOTS tools are available that enable easy creation and maintenance of ontologies for use in assessment and instruction, and we are using one of them—Protégé (Gennari et al., 2002)—to organize and represent knowledge in a form that communicates to CRESST's HPKMT authoring and scoring tools. Our long-term goal is to use the HPKMT as the primary interface to the domain expert, who uses it to express the upper-level ontology. The upper-level ontology will then be used as a roadmap for querying the domain expert for more detailed information, e.g., to elaborate on why a relationship holds, to specify concrete examples, and to differentiate the relations in terms of criticality. This querying will require a new knowledge acquisition tool extending and integrated with the HPKMT.

THREE EXAMPLES OF TESTBED RESEARCH ON PERFORMANCE ASSESSMENT IN DISTANCE LEARNING

The research described in this paper employs the HPKMT to assess a learner's understanding of a content area using graphical representation. Learners express their understanding of content by creating knowledge maps, network representations in which nodes represent concepts and links represent the relationship between two concepts. Figure 1 shows an example knowledge map for the rifle marksmanship knowledge domain created with the HPKMT. HPKMT functionality is described in detail in Chung et al. (2003). There are several knowledge mapping tools

available, but the HPKMT is one of the few designed specifically for assessment purposes and is the only system that can support multiple assessment formats. It is also the only system with an empirical base of research on its psychometric properties and quality. This section describes a research conducted in the testbed on the quality of measures obtained with the HPKMT.

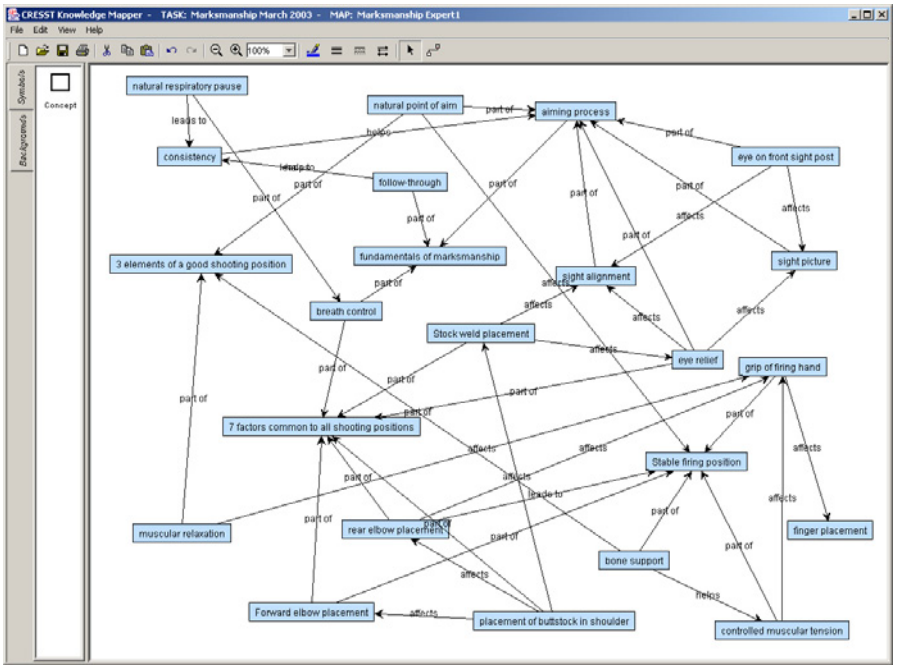

Figure 1. A knowledge map for rifle marksmanship.

Example 1: Measuring Learning Rates for Individuals and Groups

Research context. An essential determinant of quality for an assessment of learning is that it be *sensitive to instruction*. Instructional sensitivity refers to the extent to which an assessment or test detects the effects of instruction (Linn, Baker, & Dunbar, 1991; Pellegrino, Chudowsky, & Glaser, 2001). Instructional sensitivity is *not* an inherent property of assessments. For example, assessments that are intended to measure learners' progress on a broad set of standards are relatively insensitive to instruction compared to assessments that more closely sample knowledge and skills that are actually taught. A situation that commonly occurs is small changes in test scores associated with large changes instruction—leading one to wonder whether there were problems with the assessment or problems with the instruction. Thus, gathering evidence of the instructional sensitivity of a measure is important whenever the measure is used in a new or novel way (AERA, APA, & NCME, 1999; Baker, 1994; Elmore & Rothman, 1999; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002; Yoon & Resnick, 1998).

In a study reported in Kim, Chung, and Delacruz (2004), the research focused on evaluating the instructional sensitivity of knowledge mapping in the context of Marine rifle marksmanship training. The goal was to determine the degree to which knowledge mapping, repeatedly administered across the training period, could be used to measure Marines' development of rifle marksmanship knowledge.

Distance learning context. To preview the findings, the relevance of this example to distance learning is that it will show that knowledge maps are potentially useful assessments for monitoring trainees' learning throughout a training cycle. Such information, if available in real-time, could be used to provide instructors with ongoing information on the overall learning rate for the class, provide insight into individual trainee learning rates, or monitor variance from expected knowledge criteria. The critical capability is that the information be available in real-time so that timely instructional decisions can be made, and the information be actionable to support effective instruction.

Method. Fifty-three entry-level 2nd Lieutenant Marines participated in the study. In general, the Marines had little or no prior shooting experience or formal marksmanship training.

Participants were administered knowledge maps over a two-week period that spanned formal classroom training and live-fire practice. Training began midweek of the first week with two days of classroom instruction, followed by a day of practice without firing and rifle sighting. The following week consisted of live-fire practice for four days and the final day of qualification. During live-fire practice, participants could get help from range coaches. During qualification, participants did not receive any help.

Knowledge mapping. Knowledge maps were administered online to participants on six occasions: before and after classroom instruction, at the beginning and middle of live-fire practice, immediately after qualification, and four days after qualification. The mapping task required participants to modify the knowledge map from the previous occasion.

The set of concepts included *three elements of a good shooting position, seven factors common to all shooting positions, aiming process, bone support, breath control, consistency, controlled muscular tension, eye on front sight post, eye relief, finger placement, follow-through, forward elbow placement, fundamentals of marksmanship, grip of firing hand, muscular relaxation, natural point of aim, natural respiratory pause, placement of buttstock in shoulder, rear elbow placement, sight alignment, sight picture, stable firing position, stock weld placement,* and *trigger control.* The set of links described potential relationships between concepts: *affects, decreases, follows, happens during, helps, increases, leads to, part of, requires,* and *uses*.

Participant knowledge maps were scored by comparing participants' knowledge maps against criterion maps generated by three subject matter experts (primary marksmanship instructors). A participant's score was the total number of propositions in his or her map that were also in any of the criterion maps.

Analyses and results. Evidence on instructional sensitivity was gathered by examining how differences in the assessment scores related to differences in the exposure to instruction (e.g., the presence versus absence of instruction, or high versus low exposure to instruction). Because all subjects in the sample received identical instruction simultaneously, the analysis focused on how individual participants changed before and after instruction. For the knowledge mapping measure to be instructionally sensitive, there should be a significant and high improvement during the instruction period and lower or no improvement during the post-instruction period.

Figure 2 displays the observed change trajectories of knowledge map scores for the 53 participants in the study. The figure shows that many individuals improved their knowledge mapping scores (i.e., steep slopes) from pre-instruction (day 7) to post-instruction (day 9), and then remained stable after the post-instruction (i.e., flatter slopes). There is considerable individual variability in the change patterns: specifically, in terms of where they start (i.e., scores on day 7), how they change during the instruction period, and how they change during the post-instruction period. These individual differences could be due to differences in individual characteristics. For example, differences in initial status are possibly due to differences in general and content-relevant abilities following a stages-of-processing framework (Chung et al., 2004). There is less theoretical background about possible correlates of change rates (i.e., who would improve faster than others during the instructional period and after the instructional period).

Empirical analyses of learning rates. Empirical evidence of instructional sensitivity was gathered using growth modeling techniques. A key characteristic of modeling instructional sensitivity is an assumption of differential change rates for two distinct periods, the instructional and the post-instructional periods.

The results show that the "average" participant scored 12.99 in the knowledge map measure before any instruction or any live-fire experience. This finding can be interpreted as the average initial (knowledge mapping) status of a 2nd Lieutenantt with little or no marksmanship knowledge or experience.

During the classroom instruction, participants increased 1.76 points a day on average; and during the following live-fire practice, the increase is 0.37 points a day. These values can be interpreted as showing that, on average, a participant improves his or her knowledge maps at a significant and fast rate in response to the classroom instruction, and at a significant but slower rate in response to the live-fire practice. It is notable that the change rate during the first period is more than 4 times the change rate during the second period.

The significant and fast improvement during the instruction period can be interpreted as an indication of

instructional sensitivity. Presumably, for novices with little or no prior knowledge, formal instruction via classroom training would result in large changes in knowledge. The fast improvement rate observed in our sample indicates that knowledge mapping is sensitive enough to capture those rapid changes. In contrast, in live-fire practice there was no formal instruction (and presumably little learning of content), and the improvement rate observed was much lower. This decrease in rate is also evidence of instructional sensitivity, as a sensitive instrument should also detect a reduction in the rate of improvement.

Finally, the results show that there were significant variations across individual participants in initial performance, how fast performance grows in response to classroom instruction, and growth rate following classroom instruction. For example, a participant who is one standard deviation above the rate of change during classroom instruction tends to increase 4.78 points per day during the classroom instruction, while a participant who is one standard deviation below decreases 1.26 points per day during the classroom instruction.

Figure 2 also presents the expected growth trajectory of knowledge mapping scores for the population of 2nd Lieutenants who have little or no prior knowledge on rifle marksmanship when they receive a similar kind of training. The observed individual trajectories are included in the figure in order to emphasize the significant and wide individual variability in terms of where they start and how fast they learn.

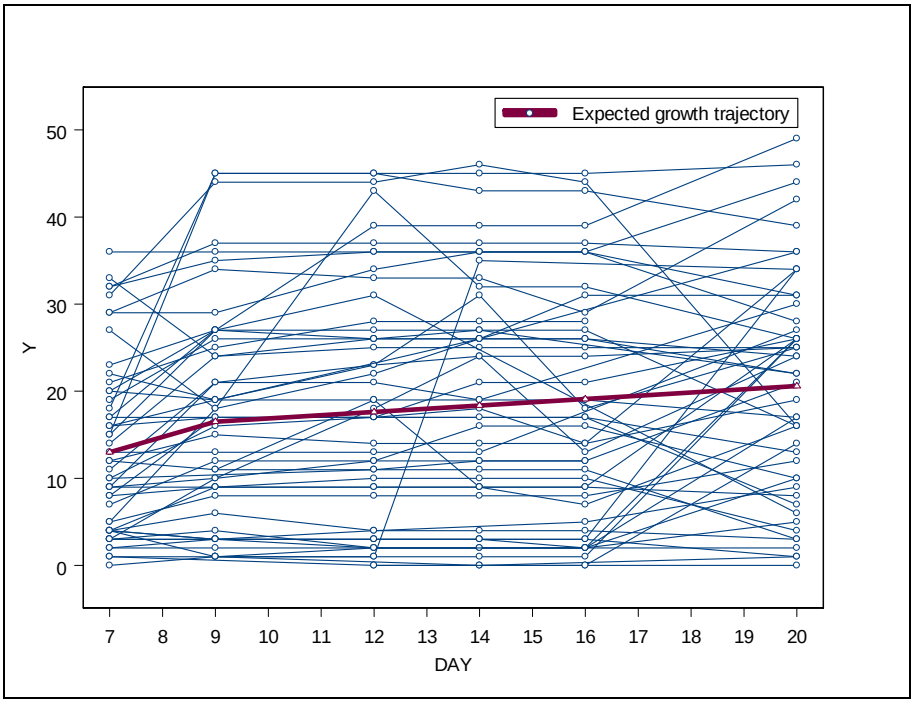

Figure 2. Observed change trajectories and the estimated change trajectory in knowledge map scores.

Discussion. One question this research addressed was the extent to which knowledge mapping can detect differences in the exposure to instruction. The results indicate that Marines' knowledge mapping scores increased at a fast rate from the pre- to post-instruction occasions, and at a significant but slower rate in follow-up occasions. This finding supports the idea that knowledge mapping is instructionally sensitive, although more work is required to confirm this finding.

In an applied setting, knowledge mapping could be used as a way to monitor trainees' learning rates via repeated and routine administrations of the knowledge map. Such an assessment method could provide deeper insight into what is learned (or not learned), when learning occurs (or does not occur), and how fast (or slow) the material is being learned. This study demonstrated that knowledge maps could be used to directly estimate improvement rates over a training day for the two different time periods. The potential for use in distance learning is that knowledge mapping could provide timely information on learning in fine or coarse detail, depending on the instructional requirements for such information.

Example 2: Automated Scoring and Validation of Constructed Responses

Research context. In a series of studies by Chung et al. (2004), the research focused on validating assessments of knowledge in the context of Marine rifle marksmanship training. A principal focus of the task was to develop ADL-compatible performance assessments—that is, online assessments with automated scoring capability that required trainees to demonstrate their understanding via assessments in a constructed-response format (vs. selected response formats such as multiple-choice).

One example of such an online assessment is a measure of shot-group knowledge. This assessment was developed to aid us in measuring one aspect of a Marine's knowledge of the fundamentals of rifle marksmanship, the relationship of shot groupings to shooting problems. Participants were required to depict shot groups associated with five shooting problems. This assessment, like the knowledge mapping of Example 1, was implemented with the HPKMT and illustrates the HPKMT's ability to support multiple assessment formats.

Distance learning context. The relevance of this example to distance learning is that it shows that scoring algorithms can be implemented for complex responses. In addition, it shows one way of validating scoring accuracy by comparing scores from the automated scoring system with scores from human raters. As with the knowledge mapping example, the utility of distance learning assessments hinges largely on whether scoring can be accomplished via automated means. Of potentially greater importance, the example provides evidence for the quality of the scoring scheme in terms of how well the automated scoring system measures knowledge.

Method: Sample. For the purpose of this example, we report data drawn from two studies involving sustainment-level Marines (N = 159 and N = 152). The main difference between the two samples is that participants in Study 1 were a little over a year younger on average than the participants in Study 2.

Shot-group depiction task. Measuring shot-group knowledge was carried out by asking participants to depict a 5-shot group for the following shooting problems: (a) firing while breathing, (b) focusing on the target (vs. focusing on the front sight post), (c) flinching, (d) bucking, and (e) improper sight adjustment. Custom-developed software was used for representation and scoring (see Chung et al., 2003 for a description of the system).

A screen shot of the computer interface is shown in Figure 3. The task required Marines to drag exactly 5 icons of each color from the left column onto the target to show the shot pattern. Each icon was colored and represented a different kind of shooter error.

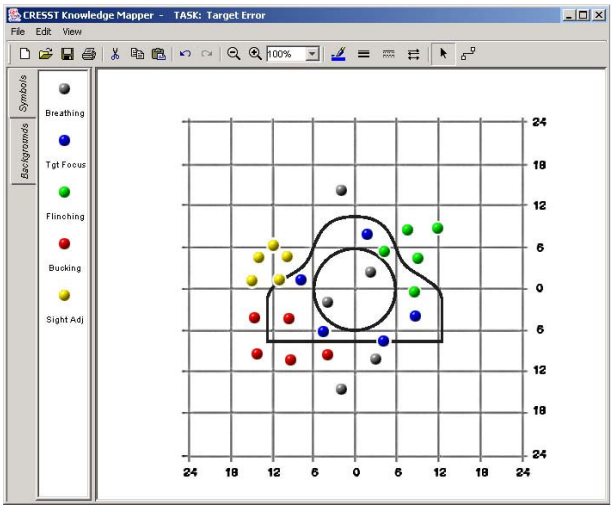

Figure 3. Sample screen shot of the shot group depiction task computer interface.

Scoring. The automated scoring algorithm was based on prior work on developing statistical measures of marksmanship (e.g., Johnson, 2001). The scoring algorithm compared the trainee's response against an expert's response. The particular variables used for the comparison differed by shot group, but in general included the center of mass of the depicted shot group and its distance from the center of the target as well as the shot group's orientation, maximum and minimum (highest and lowest shots and the horizontal and vertical ranges), the mean radius, area of dispersion, and finally the standard deviation. These measures are compared against an expert's depiction and given a binary score (match/no match).

Additional measures. Because this work was part of a large effort, additional measures were available to support validity analyses of the automated scoring system. We used the following measures: (a) background information [e.g., prior shooting experience, training experience], (b) a multiple-choice measure of basic marksmanship that sampled a broad range of topics, and (c) a computer-based measure that required participants to identify proper and improper shooting positions.

Results: Reliability of the scoring algorithm. Participants' shot group depictions were scored automatically. Each type of shot group was scored as correct or incorrect and summed to form a total score.

A scoring rubric was developed from descriptions provided by USMC marksmanship coaches and from reviews of shot group analyses in U.S. Army and USMC field manuals (U.S. Army, 1989; USMC, 1992). To establish confidence in the rubric, ratings from 2 raters were gathered and compared. Exact agreement between 2 human raters ranged from 73% to 92%.

To examine the accuracy of the automated scoring algorithm, we compared the scores from the automated system to scores from an expert rater. The analyses were based on a second sample of shot-group responses. Agreement between the automated scoring system and the expert rater was 91.2 for breathing, 88.5% for target focus and sight adjustment, 96.5% for flinching, and 94.7 for bucking.

Validity of the scoring algorithm. Data of shot group knowledge were gathered from sustainment-level Marines in two studies (N = 159 and N = 152). As noted earlier, the main difference between the two samples was that participants in Study 1 were a little over 1 year younger on average than the participants in Study 2. Participants' shot group depictions were scored automatically. Each type of shot group was scored as correct or incorrect and summed to form a total score.

In Study 1, a significant relationship was found between participants' record-fire score and their shot-group score (r = .27, p < .01). Shot-group scores also related significantly with their basic marksmanship knowledge (r = .42, p < .01), knowledge of shooting positions (r = .27, p < .01), and negatively with the number of months since their last training (r = -.27, p < .01). Additional evidence of the quality of the measure was found in Study 2. In this case, the finding of a relationship with basic marksmanship knowledge was replicated (r = .28, p < .01), and in addition, participants who had just completed a marksmanship coaches course scored significantly higher (M = 3.07, SD = 0.25) than the participants (M = 1.98, SD = 1.15) on the shot group depiction tasks (p < .001).

Discussion. As an example of automated scoring of complex performance, this study demonstrated that the measure was sensitive to presumed knowledge differences (coaches vs. non-coaches), related to other measures of knowledge (basic marksmanship knowledge, knowledge of shooting positions, and months since last training), and related to shooting performance (Study 1 only). The accuracy of the scoring algorithm was established by comparing scores from human raters with scores from the online scoring system. Validity evidence was gathered by examining the scores in the context of other knowledge measures and more knowledgeable participants.

As with the knowledge mapping assessment, the use of the online shot-grouping assessment has the potential to offer rapid feedback on performance. An important step in the use of automated systems is the establishment of first the accuracy of the scoring algorithm using standards established by expert raters, and then gathering evidence that the scores are meaningful (e.g., relate to other measures and differentiate between more and less knowledgeable participants).

Example 3: Extracting Structural Information From Knowledge Maps to Support Instruction

Research context. In an ongoing study by Lee, Chung, Cheak, Bewley, and Ellis (2004), the research question focused on the generalizability of knowledge mapping scoring. Although much research has been conducted on knowledge mapping in K-16 environments, there has been only limited work on examining the technical properties of knowledge maps in general, and virtually no work on examining the technical properties of online knowledge mapping for ADL purposes, much less in a military context. This study is intended to gather information on the reliability of online knowledge mapping in a military context. A secondary objective is to develop and validate measures of quality, derived solely from the knowledge map.

Overall, the cumulative findings reported across various empirical studies suggest that knowledge mapping is promising as a technique to measure students' knowledge of a domain. Knowledge map scores appear to differentiate between high- and low-knowledge students, to be sensitive to learning, to relate to other measures of performance, and to be sensitive to language proficiency (Chung et al., 2003).

Distance learning context. From the perspective of distance learning, scoring methods that can go beyond a single score are highly desirable (e.g., to pinpoint concepts or other information that can be acted on instructionally). Current automated scoring techniques are efficient yet do not provide guidance. Non-automated scoring methods require human raters to provide such information and are labor- and time-

intensive. Further, manual scoring results in feedback about performance that lags the instructional cycle, and thus the feedback cannot be used to effect immediate and timely changes in instruction. Thus, an automated means to provide qualitative information is highly desirable.

Method: Knowledge mapping task. We are currently collecting data. Because the focus of this research was a generalizability study concerned with consistency across tasks, participants were administered multiple knowledge maps. The domain was joint special operations in the area of (a) air tasking order (ATO) cycle, (2) joint task force (JTF) structures and functions, and (3) joint special operations task force (JSOTF) structure. For the purposes of this paper, the focus will be on the ATO cycle.

Four experts determined by consensus the relevant concepts and relationships for ATO. Each expert also created a knowledge map to be used for scoring purposes. Thirty-two concepts were include and 14 links. The set of 32 concepts included *ACMREQs, ACO, AIRSUPREQs, ALLOREQs, ATO, ATO development, ATO execution, Airspace Requests, Apportionment, Approved JIPTL, Draft JIPTL, Assignment to unit, Battle damage assessment, Close air support, Combat assessment, Commander's intent, Component input, Interdiction, JFC Guidance component coordination, JTCB, JGAT, MAAP, Restricted target list, SOF forces, Sortie generation, Strategic attack, Support, Target development, TBMCS, Weaponeering allocation,* and *Weaponization of targets*. The set of links were *derived from, followed by, input, leads to, output, produces,* and *supports*.

Scoring. One aspect of this research was determining how to extract assessment information from knowledge maps that could (a) be used for instructional purposes, (b) be derived directly from the representation via software, and (c) complement our existing scoring technique.

We developed a technique that analyzes knowledge maps with respect to their structure or connectivity. The focus of the scoring is on the way knowledge is organized, in terms of the network of nodes and their interconnectedness. The purpose behind this type of scoring is to identify patterns in the knowledge space. The components of the structural scoring include the following properties that can be used to characterize different aspects of a knowledge map:

Number of nodes used. This measure is a count of the number of nodes used in a student's knowledge map.

Type of node. A node can be characterized as a source, sink, or carrier. A source is a node with only outgoing links, a sink is a node with only incoming links, and a carrier has both types of links. Associated with type of node is the number of incoming links (fan-in) and outgoing links (fan-out).

Node clusters. A cluster is a set of connected concepts and is an important feature of map organization because it helps differentiate concepts into key concepts and supporting ones. A node within a cluster that has a high number of fan-in or fan-out connections can be considered a focal point and thus a target of instruction.

Reachability. This measure is the number of other nodes reachable from a given node. High reachability indicates that the network is highly interconnected. Low reacheability indicates that the network is sparse and linear.

Results. A preliminary analysis of the data collected thus far (N = 29) suggests that the structural measures capture differences between expert maps and student maps, and differences among students' maps relative to structural complexity. In general, the expert maps had more terms, variable use of sources, sinks and carriers, numerous clusters, and high reachability. Additionally, a comparison of a sample of student maps revealed similar patterns, with more sophisticated maps containing a higher number of terms, links, and clusters as well as level of integration (reachability). Among expert maps, the following key terms around which clusters occur were *Apportionment, ATO development, ATO execution, Combat assessment, Target development,* and *MAAP*.

Discussion. One of the objects of the structural analyses is to locate clusters and thus provide a means to compare student and expert maps. Student maps lacking clusters that are in expert maps may be one way to identify areas of conceptual weakness. For example, if a student map is missing a key concept, or if the key terms are missing or poorly elaborated by supporting terms, the identification of the missing clusters can be the basis for remediation.

SUMMARY AND DISCUSSION

In this paper we presented our current distance learning testbed implementation and examples of research conducted in the testbed on methods for assessing learner performance in a distance learning environment. We described seven major testbed

components: data collection platforms, special data collection tools, data analysis and reporting tools, a courseware rating tool, assessment authoring tools, performance assessment scoring tools, and knowledge acquisition and representation tools. We also illustrated, through three examples, how the tools have supported our current research efforts.

A common thread in the design of the testbed tools is the idea that the tools should support and be supported by empirical research. That is, on the one hand the testbed functions as infrastructure to support basic research on issues related to assessment and learning. On the other hand, we intentionally design the tools to be relevant to operational settings so that if promising research findings emerge, the tools can be moved quickly to an operational mode to support solving real-world problems.

A second theme is the continual validation (or revalidation) of the tools for the various purposes to which they are applied. For example, knowledge mapping has been validated in public school settings where education is the goal and instruction is delivered in formal and structured settings. Far less is known about how knowledge mapping would operate in a military training setting where, for example, the objective is to develop knowledge related to skill acquisition such as rifle marksmanship.

Finally, a fundamental aspect of the testbed is to investigate novel ways of assessing human performance—domain knowledge, problem solving, and teamwork, for example—in the context of computer-based assessments in general. To this end, the critical starting point is the cognitive demands of the task—articulating the set of cognitive processes required for success in the target domain first (e.g., understanding of fundamentals of rifle marksmanship), and then designing methods of measuring these processes and learning outcomes. An important part of this process is operationalizing the scoring algorithm for measuring "understanding" or "problem solving" and then validating the approach with empirical evidence.

As the Armed Services turn increasingly to ADL to deliver training and education, there is the expectation that ADL systems can and will deliver quality training—to the right people, at the right time, and at the right place—to support operational readiness and personal excellence (e.g., Air Force Institute for Advanced Distributed Learning, 2001; Department of Defense, 1999; Director of Naval Training [N7], 1998; U.S. Army Training and Doctrine Command, 1999). The development of the technical infrastructure and standards is currently underway (e.g., Advanced Distributed Learning [ADL], 2003a) as well as guidelines for effective ADL implementation (ADL, 2003b). However, most assessment tools only provide methods to measure recall and recognition, and these formats may not provide enough information to answer questions about ADL training effectiveness on complex tasks. Our development of a testbed is one way to move toward scalable methods for measuring complex human performance.

ACKNOWLEDGEMENTS

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Educational Science, U.S. Department of Education, and the Office of Naval Research Award Number N00014-02-1-0179, as administered by the Office of Naval Research. The findings and opinions expressed in this report do not reflect the positions or policies of the U.S. Department of Education, or the Office of Naval Research.

REFERENCES

Advanced Distributed Learning. (2003a). *Advanced Distributed Learning Sharable Content Object Reference Model Version 1.2* [On-line]. Available: http://www.adlnet.org

Advanced Distributed Learning. (2003b). *What works in distance learning*. H. F. O'Neil (Ed.). [On-line]. Available: http://www.adlnet.org

Air Force Institute for Advanced Distributed Learning. (2001). *Air Force Advanced Distributed Learning Vision*. Maxwell AFB-Gunter Annex, AL: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Army Science Board. (1997). *Distance learning.* Arlington, VA.

Baker, E. L. (1994, March). Making performance assessment work: The road ahead. *Educational Leadership*, *51*(6), 58-62.

Baker, E. L. (Ed.). (1996). A focus on educational assessment [Special issue]. *Journal of Educational Research*, *89*(4).

Baker, E. L., & Herman, J. L. (2003). A distributed evaluation model. In G. D. Haertel & B. Means (Eds.), *Evaluating educational technology*. New York: Teacher's College Press.

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131-153). Englewood Cliffs, NJ: Prentice-Hall.

Baker, E. L., Niemi, D., Herl, H., Aguirre-Muńoz, A., Staley, L., & Linn, R. L. (1996). *Report on the content area performance assessments (CAPA): A collaboration among the Hawaii Department of Education, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the teachers and children of Hawaii* (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Barry, M., & Runyan, G. B. (1995). A review of distance-learning studies in the U.S. Military. *American Journal of Distance Education, 9,* 37–47.

Chung, G. K. W. K., Baker, E. L., Brill, D. G., Sinha, R., Saadat, F., & Bewley, W. L. (2003, December). Automated assessment of domain knowledge with online knowledge mapping. *Proceedings of the I/ITSEC*, Orlando, FL.

Chung, G. K. W. K., Delacruz, G. C., de Vries, L. F., Kim, J.-O., Bewley, W. L., de Souza e Silva, A. A., et al. (2004). *Determinants of rifle marksmanship performance: Predicting shooting performance with advanced distributed learning assessments* (Deliverable to Office of Naval Research). Los Angeles: University of California, CRESST.

Chung, G. K. W. K., O'Neil, H.F., Jr., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior, 15,* 463-493.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*, 445–459.

Clark, R. E. (1989). Current progress and future directions for research in instructional technology. *Educational Technology Research & Development, 37*, 57–66.

Clark, R. E. (Ed.). (2001). *Learning from media: Arguments, analysis, and evidence*. Greenwich, CT: Information Age Publishing.

Delacruz, G. C., Chung, G. K. W. K., & Bewley, W. L. (2003, December). Identifying learning phases using the Human Performance Knowledge Mapping Tool (HPKMT) and microgenetic analysis. *Proceedings of the I/ITSEC*, Orlando, FL.

Department of Defense. (1999). *The Department of Defense Advanced Distributed Learning Strategic Plan.* Washington, DC: Pentagon.

Director of Naval Training (N7). (1998). The *Navy-wide distributed learning planning strategy.* Washington, DC: Navy Pentagon.

Elmore, R. F., & Rothman, R. (Eds.). (1999). *Testing, teaching, and learning: A guide for states and school districts.* Washington, DC: National Research Council.

Fensel, D., Hendler, J., Lieberman, H., & Wahlster, W. (Eds.). (2003). *The semantic web: Why, what, and how*. Cambridge, MA: MIT Press.

Gennari, J., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubézy, M., Eriksson, H., et al. (2002). *The evolution of Protégé: An environment for knowledge-based systems development* (Stanford Medical Institute Tech. Rep. No. 2002-0943). Stanford University: Palo Alto, CA.

Herl, H. E., Baker, E. L., & Niemi, D. (1996). Construct validation of an approach to modeling cognitive structure of U.S. history knowledge. *Journal of Educational Research, 89,* 206-218.

Herl, H. E., O'Neil, H. F., Jr., Chung, G. K. W. K., & Schacter, J. (1999). Reliability and validity of a computer-based knowledge mapping system to measure content understanding. *Computers in Human Behavior, 15*, 315–334.

Johnson, R. F. (2001). Sta*tistical measures of*

marksmanship (ARI Tech. Note TN-01/2). Alexandria, VA: U.S. Army Research Institute.

Kim, J-O., Chung, G. K. W. K., & Delacruz, G. C. (2004, April). *Examining the sensitivity of knowledge maps using repeated measures: A growth modeling approach.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Lee, J. J., Chung, G. K. W. K., Cheak, A. C., Bewley, W. L., & Ellis, K. (2004). *An investigation of the reliability of knowledge measures through relational mapping in joint military environments*. Unpublished manuscript.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

Lockee, B. B., Burton, J. K., & Cross, L. H. (1999). No comparison: Distance education finds a new use for "no significant difference." *Educational Technology Research & Development, 47*, 33–42.

Machtmes, K., & Asher, J. W. (2000). A meta-analysis of the effectiveness of telecourses in distance education. *American Journal of Distance Education, 12*, 7–25.

Madden, J. (1998). *Issues involved with successful implementation of distributed learning* (A report to the Office of Training Technology). Orlando, FL: Naval Air Warfare Center.

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessments* (Committee on the Foundations of Assessment; Board on Testing and Assessment, Center for Education. Division on Behavioral and Social Sciences and Education, National Research Council). Washington, DC: National Academy Press.

Phipps, R., & Merisotis, J. (1999). *What's the difference? A review of contemporary research on the effectiveness of distance learning in higher education*. Washington, DC: Institute for Higher Education Policy.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research on Science Teaching, 39*, 369–393.

Russell, T. L. (1999). *The No Significant Difference phenomenon* (0-9668936-0-3). North Carolina: NC State University.

Saba, F. (2000). Research in distance education: A status report. *International Review of Research in Open and Distance Learning, 1*. [On-line]. Available: www.irrodl.org

Schacter, J., Herl, H. E., Chung, G. K. W. K., Dennis, R. A., & O'Neil, H. F., Jr. (1999). Computer-based performance assessments: A solution to the narrow measurement and reporting of problem solving. *Computers in Human Behavior, 15*, 403-418.

Smith, P. L., & Dillon, C. L. (1999). Comparing distance learning and classroom learning: Conceptual considerations. *American Journal of Distance Education, 13*, 6–36.

U.S. Army. (1989). *M16A1 and M16A2 Rifle Marksmanship* (FM 23-9). Fort Benning, GA: Author.

U.S. Army Training and Doctrine Command. (1999). *The Army Distance Learning Plan*. Fort Monroe, VA: Author.

U.S. Marine Corps. (1992). *Basic marksmanship* (PCN 139 000023 00, FMFM 0-8). Quantico, VA: Author.

Wisher, R. A., Champagne, M. V., Pawluk, J. L., Eaton, A., Thornton, D. M., & Curnow, C. K. (1999). *Training through distance learning: An assessment of research findings* (Tech. Rep. No. 1095). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Yoon, B., & Resnick, L. B. (1998). *Instructional validity, opportunity to learn, and equity: New standards examinations for the California mathematics renaissance* (CSE Tech. Rep. No. 484). Los Angeles: University of California, CRESST.